DOCUMENT RESUME

ED 465 785 TM 034 170

AUTHOR Bastick, Tony
TITLE Demonstrating Local Item Dependence for Recognition and Supply Format Tests.
PUB DATE 2002-04-00
NOTE 9p.; Paper presented at the Annual Meeting of the Western Psychological Association (Irvine, CA, April 14-17, 2002).
PUB TYPE Reports - Research (143) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Constructed Response; *Elementary School Students; Elementary Secondary Education; Foreign Countries; Multiple Choice Tests; Performance Factors; *Secondary School Students; *Test Construction; Test Format; Test Items
IDENTIFIERS *Item Dependence

ABSTRACT

This study tested the hypothesis that the common approach to test construction in which recognition questions (RQs), such as multiple-choice items, are followed by constructed response questions (CRQs) encourages students to use the informationally rich RQs to gain marks on the CRQs, thus introducing Local Item Dependence (LID) and inflating the CRQ test scores. This was tested with 188 children aged 10 to 16 years in 5 schools using class tests in 4 topic areas. The children in each class were randomly assigned to take the test in the traditional RQ-CRQ order, or in the experimental CRQ-RQ order. Using two independent t-tests, the groups were then compared on their RQ scores and on their CRQ scores. The results indicate that a statistically significant advantage was gained on the CRQs when the traditional order of test construction was used. Differences in mean RQ scores were used to check if factors other than LID, which could be associated with the nontraditional order, might have influenced CRQ results. These checks showed no statistically significant differences between the two groups. It is concluded that the traditional order can produce LID and result in inflated test scores for the constructed response part of the test. (Contains 1 figure, 1 table, and 24 references.) (SLD)

Demonstrating Local Item Dependence for Recognition and Supply Format Tests

Tony Bastick
University of the West Indies

Paper presented at the Western Psychological Association
WPA 2002 Convention
Irvine, CA
USA

Publication Date: April, 2002

TM

Demonstrating Local Item Dependence for Recognition and Supply Format Tests

Tony Bastick
University of the West Indies

Abstract

Ideally, each question on a test should be an independent sample of the testee's ability. When this is not the case, the information from one question can be used to gain marks on another question. This is an example of Local Item Dependence (LID) and its occurrence can inflate resulting test scores.

A common form of test construction is one where recognition questions (RQs), such as Multiple-choice items are followed by constructed response questions (CRQs) such as short answer items. This research hypothesised that this common construction encourages students to use the informationally rich RQs to gain marks on the CRQs, thus introducing LID and so inflating the CRQ test scores.

This was tested with children (n=188, age 10-16 yrs) in five schools using class tests in four topic areas. The children in each class were randomly assigned to take the test in the traditional RQ-CRQ order or in the experimental CRQ-RQ order. Using two independent t-tests, the groups were then compared on their RQ scores and on their CRQ scores. The results indicated that a statistically significant advantage was gained on the CRQs when using the traditional order of test construction. Differences in mean RQ scores were used to check if factors other than LID, which could be associated with the non-traditional order, might have influenced CRQ results. These checks showed no statistically significant differences between the two groups.

It was therefore concluded that the traditional order of recognition questions followed by constructed response questions can produce Local Item Dependence and result in inflated test scores for the constructed response part of the test.

Introduction

A basic assumption of an objective test, and a psychometric requirement, is that the questions independently sample the test-taker's abilities. However, tests commonly use both recognition questions (RQs) such as multiple choice questions and constructed response questions (CRQs) such as short answer questions. Further, the RQs are usually presented first, probably because it is felt that students can better pace their responses and that the easier response format of ticking a box or circling an option allows for more questions to be completed, and hence a better sampling of ability, than the more time consuming requirement to write an answer. Because there is more content information in RQs than in CRQs, this standard sequence of RQs followed by CRQs makes it possible that the testee transfers information supplied in the RQs e.g. in the multiple-choice stem and choices, to cue answers to the CRQs. If this is so then the traditional sequence of questions contradicts the basic test assumption that questions are independent and gives inflated results for the second part of the test. It is important to test this possibility because, if it is so, the widespread use of this sequencing implies that many test results are inflated in comparison to single genre tests.

Bastick, T. (2002, April). *Demonstrating Local Item Dependence for Recognition and Supply Format Tests.* Paper presented at the Western Psychological Association, WPA 2002 Convention, Irvine, CA. USA.

Many examinations use both recognition items and constructed response items

Examinations throughout the world commonly use a combination of recognition and constructed response questions. Examinations in most subjects for University examinations and many teacher-made classroom tests ask both recognition and constructed response questions. Combining both genres of questions is also a common practice in high-stakes examinations for many subjects for the GCSE in the UK, for the CXC Caribbean examinations, and for the Advanced Placement examinations in the USA (Lukhele, Thissen & Wainer, 1994).

Both RQs and CRQs, in their different formats, are used because of the different advantages they offer (Gallagher, 1998, pp. 130-132; Linn & Gronlund, 2000, pp. 236-242; Mehrens & Lehmann, 1991, pp. 66-67; Shepard, 1996).

Effects of combining the two genres in the same examination

Given the widespread practice of combining both RQs and CRQs on the same examination, it is important that we understand the influences these genres have on one another and on how this impacts on students' performance. For example, does the order in which students answer these questions affect their results?

Multiple-choice and short answer questions are the most common item formats on combined genre tests (New Hampshire State Department of Education, 1994) and are the most common to be analysed together. Research findings have indicated that although the results of MC and CRQs tend to be correlated (Martinez, 1990; Pollock, 1997), it is usual that scores on the MC sections are higher than on the CRQs (Bay, 1998; DeMars, 1998; Dossey, 1993).

Studies that have compared the scores of males and females on RQ and CRQ tests have found that differences are linked to content and item format (Garner, & Engelhard 1999; Pomplun, & Sundbye, 1999). For example, it has been found that in science males do better on CRQs involving visualizations and CRQs that call upon knowledge and experience acquired outside of school (Hamilton, & Snow, 1998) and in mathematics males score higher on CR problem solving questions (Wilson, & Zhang, 1998). Although the above studies reported gender differences linked to content and item format, Christine DeMars (1998) compared males and females in mathematics and science on MC and CRQs in 201 schools and found no overall gender difference. Larger differential MC and CR scores have been found for younger students. Tahany Gadalla (1999) tested stem-equivalent and scoring-equivalent MC and CR forms of the Canadian Achievement Tests with 1,028 students in grades 2 through 6 and found that differences between MC and CR scores were greater in the grades 2 and 3 as compared to grades 4, 5, and 6.

MC and CRQ performance differences have also been compared across ethnic and Social Economic Status groups. James Myerberg, (1996) compared students from grades 3 to 8 in mathematics, language arts and reading on the Maryland School Performance Assessment program and found CRQs favoured female students, non-white students and students from low SES backgrounds. Students from fifth grade to high school level have also been found to have great variability in their affective responses to the different genres (Hamilton, 1994) and reported to have greater confidence for CRQs (Barnett-Foster, & Nagy, 1995).

Expectations that different formats measure different abilities

There is an increasing trend towards high-stakes examinations that combine multiple-choice items with open-ended tasks, particularly for Teacher Certification (Klein, 1998). The fact that educators persist in using constructed response items rather than more easily marked recognition items, even in the face of the considerable extra cost in time required to mark them, indicates that there is an expectation that constructed response questions are perhaps adding a different dimension to the examination of students. A study of RQ and CRQs by Ercikan, Schwarz and others (1998) gave some support to this view by reporting a factor for each genre, although the factors were highly correlated. In-line with this belief, the University of the West Indies Examination Regulation 28 (iv), actually restricts the use of RQs to a maximum of only 25% of student assessment (University of the West Indies, 2000, p. 8).

Purpose and rationale of this study

It is possible that the traditional RQ – CRQ sequence encourages a violation of the independence assumptions of testing by allowing students to use the information rich first part of the test to answer the CRQs in the second part of the test. Although mixed genre tests have been analysed with respect to comparative difficulty, gender, subject, ethnicity and age, there does not seem to be any published research testing whether the traditional order violates the independence assumption of the test questions.

This possibility was tested by dividing test takers into two groups and giving a mixed genre test in two different orders to both groups. The scores of the group that did the RQs first were then compared to the scores of the group who did the CRQs first to find if the test order conferred any advantage. It seems probable that students will use content information gleaned from test questions to help them answer other questions thus violating the requirement of local independence. Hence, it was expected that the students who took the questions in the traditional order would be at an advantage.

Design of the experiment

Subjects

The following experiment was replicated in 5 classes in 5 different secondary schools across 4 topics with n=188 students consisting of 65 boys and 123 girls aged from 10 to 16 years with a mean age of 12 years 8 months. The numbers of males and females tested in each of the topics were Biology 1 – 'sexual reproduction in flowering plants' (m=13, f=16, n=29), Biology 2 – 'Endocrine Systems' (m=18, f=15, n=42), Physics – 'States of matter' (m=18, f=15, n=33), English – 'Nouns' (m=15, f=35, n=50) and Social Studies - 'The Family and The Peer Group' (m=11, f=23, n=34). The sample of schools, whose principals agreed to their schools and teachers participating in this research, were drawn from high and low socio-economic-status populations and represented both urban and rural areas in and around Kingston, Jamaica, West Indies.

A 45-minute in-class test was given to assess one of the above topics that had recently been taught to the class. The test was composed of six subtests that were judged by three teachers to be of equivalent difficulty. The three teachers comprised two researchers and the teacher who had recently taught the topic to the class. The six subtests each contained six questions. Three subtests were all recognition type questions - multiple-choice questions, true/false questions and matching questions. These were placed on side A of the test paper. The other three subtests were all construction questions – direct questions, completion questions and association questions - and this were placed together on side B of the paper. The questions on each side of the test paper were obviously different for each topic and randomised differently for each topic. Figure 1 illustrates the design of the test sheets.

Figure 1: Design of test sheets

Side A of test sheet	18 Randomised Recognition Questions		
Format of questions	Six Multiple Choice Questions	Six True/False Questions	Six Matching Questions

Side B of test sheet	18 Randomised Constructed Response Questions		
Format of questions	Six Direct Questions	Six Completion Questions	Six Association Questions

Four open-ended 'filler' questions were printed at the bottom of each side to reduce possible disruption by early finishers. These were "Which question was the most difficult?" and "Why?" Also "Which question was the easiest?" and "Why?" Side A carried questions on the testee to enable feedback on results to be returned to each class teacher and each side had an 'order' box at the top of the sheet.

A typical RQ and CRQ from a Physics test on 'States of matter' was:

RQ: Alcohol bubbles at 78 C when heated. At 100 C it would have

A. melted B. evaporated C. sublimed D. condensed

CRQ: All material in our world exists in _____ states.

Random assignment

All students in each class were randomly assigned to one of two groups that were approximately equal in number, plus-or-minus one. One group was the control group who were to answer Side A first, which was in the traditional order, that is RQs first followed by the CRQs. The other half were the experimental group who were to answer Side B first that is, in the experimental order of CRQs followed by RQs.

Administration

Twenty minutes were allotted to side A of the test and twenty-five minutes to Side B. The extra five minutes was to allow for the extra time needed to physically write the constructed responses as opposed to the shorter time needed to indicate recognition by using a tick, line or circle for Side A. Students all sat with Side A facing upwards ready to complete the testee information on the top of Side A. Each half was then randomly assigned, by a spin of a coin, to complete Side A or Side B first and sat with the appropriate side of the paper facing upwards. The supervisor(s) checked this. Students were told that they would all have the same time of 45 minutes to complete the test, 20 minutes for Side A and 25 minutes for Side B. There was an 'order' box at the top of each side. All students then wrote '1' in this box on the side that was facing upwards, that is the side they would be attempting first.. This was to ensure that if the papers from the control and experimental groups were later inadvertently mixed they could be correctly re-sorted. Students were informed that they would be told when to stop and must not turn over to the other side unless told to do so. The test then started. After 20 minutes those completing Side A were told to stop and turn over, write '2' in the order box and complete Side B. After another 5 minutes, those completing Side B first were told to stop, turn over, write '2' in the order box and complete Side A. During the test, supervisor(s) noted their observations of students' test taking

behaviours. After a further 20 minutes, 45 minutes in all from the start of the test, all students were told to stop and the supervisor(s) collected the papers separately from the control group and from the experimental groups. Afterwards, students were informally interviewed about their reactions to the test.

Results

T-tests were calculated on the difference between the mean scores of the two groups for the recognition questions and for the constructed response questions. The maximum score on the 18 questions of each type was 18, being one mark per question. The results for the 188 students are presented in Table 1.

Table 1: Comparisons of mean scores for the control and experimental groups

	Mean Scores	
n=188	Recognition Questions	Constructed Questions
Order		
Recognition 1st	11.0101 (61%)	6.5960 (37%)
Constructed 1st	10.6854 (59%)	5.7303 (32%)
% Difference	0.3247 (2%)	0.8657 (5%)
Significance	0.458	0.043

Table 1 shows the significance of the differences in mean scores and the percentage advantages of the traditional sequence of recognition questions first.

As shown in Table 1, students in the control group who completed the recognition questions first had slightly higher scores on the recognition items and significantly higher scores on the Constructed response questions than did students in the experimental group. A 5% increase on the 32% score of students who did the Constructed Response Questions first is a significant 16% advantage for the 'Recognition first' test order. To better understand these results, they need to be considered in conjunction with class observations, post-test interview data as presented in the following discussion.

Discussion

The question this experiment was designed to answer was: 'Does the test order of Recognition Questions first or Constructed Response Questions first, give any advantage?' We found that the RQ – CRQ order gave an advantage on both types of questions and significant advantage on CRQs. If this advantage is due to cognitive transfer as we expected, then the traditional order does violate the assumptions of item independence and is inflating the scores on the second part of the test.

From their notes of the students' test taking behaviours and their subsequent interviews, the test administrators reported that the students randomly assigned to Side A first, the Recognition questions, were observed to settle quickly into focused responding, whereas the students randomly assigned to start Side B first, the constructed Response Questions, were more agitated and asked more questions of the supervisors. In post-test interviews students said that they preferred the easier response mode of ticking and circling for the Recognition questions. It seems from these observations that the traditional RQ-CRQ order creates lower stress. When we consider that the RQs and CRQs were written to be of equal difficulty, this lower stress is likely to be due to the easier form of responding. This could be tested by replicating the experiment with groups of subjects who have different levels of language

ability. If lower stress had given some advantage, then it would have been expected to be apparent in different mean scores on the RQs. However, as there was no significant difference between the control and experimental group scores on the RQs, it can be assumed that any advantages of lower stress were also not significant for the CRQs and that the different scores are in large part due to transfer of cognition from the first part to the second part of the test that is, in the traditional sequence of recognition questions followed by constructed response questions, students are learning from the test.

References

Barnett-Foster, D. & Nagy, P. (1995). A Comparison of Undergraduate Test Response Strategies for Multiple-Choice and Constructed-Response Questions. *Alberta Journal of Educational Research, (41)*1, 18-35.

Bay, L. (1998, April). *Comparing Student Performance on Different Item Formats Relative to Achievement Levels Cutpoints.* Paper presented at the Annual Meeting of the National Council on Measurement in Education, San Diego, CA, USA.

DeMars, C. (1998, April). *Item Estimates under Low-Stakes Conditions: How Should Omits Be Treated?* Paper presented at a poster session at the Annual Meeting of the American Educational Research Association, San Diego, CA, USA.

Dossey, J. A. (1993). *Can Students Do Mathematical Problem Solving? Results from Constructed-Response Questions in NAEP's 1992 Mathematics Assessment.* (Report No. ETS-R-23-FR01). Educational Testing Service, Princeton, NJ. USA.

Ercikan, K., Schwartz, R. D., Julian, M. W., Burket, G. R., Weber, M. M., & Link, V. (1998). Calibration and Scoring of Tests with Multiple-Choice and Constructed-Response Item Types. *Journal of Educational Measurement, (35)*2, 137-54.

Ferrara, Steven (1995, April). *A Beginning Validation of Causes of Local Item Dependence in a Large Scale Hands-On Science Performance Assessment.* Paper presented at the Annual Meeting of the National Council on Measurement in Education, San Francisco, CA, USA.

Gadalla, T. M. (1999, April). *Multiple-Choice versus Constructed-Response Tests in the Assessment of Mathematics Computation Skills.* Paper presented at the Annual Meeting of the American Educational Research Association, Montreal, Quebec, Canada.

Gallagher, J. D. (1998). *Classroom assessment for teachers*. Columbus, Ohio, USA: Merrill.

Garner, M. & Engelhard, G. Jr., (1999). Gender Differences in Performance on Multiple-Choice and Constructed Response Mathematics Items. *Applied Measurement in Education, 12*(1), 29-51.

Hamilton, L. S. & Snow, R. E. (1998). *Exploring Differential Item Functioning on Science Achievement Tests.* (CSE Technical Report 483). National Center for Research on Evaluation, Standards, and Student Testing, Los Angeles, CA. USA.

Hamilton, L. S. (1994, April). *An Investigation of Students' Affective Responses to Alternative Assessment Formats.* Paper presented at the Annual Meeting of the National Council on Measurement in Education, New Orleans, LA, USA.

Klein, S. P. (1998). Standards for Teacher Tests. *Journal of Personnel Evaluation in Education, 12*(2), 123-138.

Linn, J. E. & Gronlund, M. A. (2000). *Measurement and assessment in teaching* (8th ed). Columbus, Ohio, USA: Merrill.

Lukhele, R., Thissen, D, & Wainer, H. (1994). On the Relative Value of Multiple-Choice, Constructed Response, and Examinee-Selected Items on Two Achievement Tests. *Journal of Educational Measurement, 31*(3), 234-50.

Martinez, M. E. (1990). *A Comparison of Multiple-Choice and Constructed Figural Response Items.* (Report No. ETS-RR-90-19). Educational Testing Service, Princeton, N.J. USA.

Mehrens, W. A. & Lehmann, I. J. (1991). *Measurement and evaluation in Education and Psychology* (4th ed). Orlando, Florida: Harcourt Brace Jovanovich.

Myerberg, N. J. (1996, April). *Performance on Different Test Types by Racial/Ethnic Group and Gender.* Paper presented at the Annual Meeting of the American Educational Research Association, New York, NY, USA.

New Hampshire State Dept. of Education (1994). *New Hampshire Educational Assessment Report, Grade 3.* New Hampshire State Dept. of Education, Concord: Author.

Pollock, J. M. (1997). *Constructed Response Tests in the NELS:88 High School Effectiveness Study. National Education Longitudinal Study of 1988 Second Followup. Statistical Analysis Report.* (Report No. NCES-97-804). National Center for Education Statistics (ED), Washington, DC. USA.

Pomplun, M. & Sundbye, N. (1999). Gender Differences in Constructed Response Reading Items. *Applied Measurement in Education, 12*1(1), 95-109.

Shepard, L. A. (1996, September). *Measuring Achievement: What Does It Mean To Test for Robust Understanding? William H. Angoff Memorial Lecture Series.* Paper presented at the 3rd annual William H. Angoff Memorial Lecture Educational Testing Service, Princeton, NJ, USA

Traub, R. E. (1993). On the equivalence of the traits assessed by multiple-choice and constructed-response tests. In R. E. Bennett & W. C. Ward (Eds.), *Construction versus choice in cognitive measurement: Issues in constructed response, performance testing, and portfolio assessment.* Hillsdale, NJ: Erlbaum.

University of the West Indies (2000). *Examination regulations for first degrees, diplomas and certificates 2000/2001.* Kingston, Jamaica: Author.

Wilson, L. D. & Zhang, L. (1998, April). *A Cognitive Analysis of Gender Differences on Constructed-Response and Multiple-Choice Assessments in Mathematics.* Paper presented at the Annual Meeting of the American Educational Research Association, San Diego, CA, USA.

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

TM034170

REPRODUCTION RELEASE

(Specific Document)

I. DOCUMENT IDENTIFICATION:

Title:	Demonstrating Local Item Dependence for Recognition and Supply Format Tests.
Author(s):	Bastick, Tony
Corporate Source: Paper presented at the Western Psychological Association, WPA 2002 Convention, Irvine, CA. USA.	Publication Date: 2002, April

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic media, and sold through the ERIC Document Reproduction Service (EDRS). Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce and disseminate the identified document, please CHECK ONE of the following three options and sign at the bottom of the page.

The sample sticker shown below will be affixed to all Level 1 documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

1

Level 1 ☑

Check here for Level 1 release, permitting reproduction and dissemination in microfiche or other ERIC archival media (e.g., electronic) *and* paper copy.

The sample sticker shown below will be affixed to all Level 2A documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE AND IN ELECTRONIC MEDIA FOR ERIC COLLECTION SUBSCRIBERS ONLY HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

2A

Level 2A ☐

Check here for Level 2A release, permitting reproduction and dissemination in microfiche and in electronic media for ERIC archival collection subscribers only

The sample sticker shown below will be affixed to all Level 2B documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

2B

Level 2B ☐

Check here for Level 2B release, permitting reproduction and dissemination in microfiche only

Documents will be processed as indicated provided reproduction quality permits.
If permission to reproduce is granted, but no box is checked, documents will be processed at Level 1.

I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce and disseminate this document as indicated above. Reproduction from the ERIC microfiche or electronic media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries.

Sign here, → please

Signature:	Printed Name/Position/Title: Tony Bastick, Research Coordinator, Dr.	
Organization/Address: University of the West Indies, Department of Educational Studies, Mona Campus, Kingston 7, Jamaica	Telephone: (876)927-2130	FAX: (876)977-0482
	E-Mail Address: tbastick@uwimona.edu.jm	Date: 4th April 2002

(over)

III. DOCUMENT AVAILABILITY INFORMATION (FROM NON-ERIC SOURCE):

If permission to reproduce is not granted to ERIC, *or*, if you wish ERIC to cite the availability of the document from another source, please provide the following information regarding the availability of the document. (ERIC will not announce a document unless it is publicly available, and a dependable source can be specified. Contributors should also be aware that ERIC selection criteria are significantly more stringent for documents that cannot be made available through EDRS.)

Publisher/Distributor:
Address:
Price:

IV. REFERRAL OF ERIC TO COPYRIGHT/REPRODUCTION RIGHTS HOLDER:

If the right to grant this reproduction release is held by someone other than the addressee, please provide the appropriate name and address:

Name:
Address:

V. WHERE TO SEND THIS FORM:

Send this form to the following ERIC Clearinghouse:

However, if solicited by the ERIC Facility, or if making an unsolicited contribution to ERIC, return this form (and the document being contributed) to:

ERIC Processing and Reference Facility
4483-A Forbes Boulevard
Lanham, Maryland 20706

Telephone: 301-552-4200
Toll Free: 800-799-3742
FAX: 301-552-4700
e-mail: ericfac@inet.ed.gov
WWW: http://ericfac.piccard.csc.com